Distributed Data Access, Analysis and Standards for Earth Science Data

A White Paper for the Strategic Evolution of ESE Data Systems (SEEDS) Public Workshop
February 5-8, 2002

Menas Kafatos[1]
Center for Earth Observing and Space Research (CEOSR)
George Mason University
E-mail: mkafatos@gmu.edu
[1]On behalf of the SIESIP Team

Introduction

A typical question being asked by Earth scientists today is, what observational or model output data sets exist that can be used to address my specific scientific problem? This question arises because of more open Earth science data policies, the rapid growth of the number of data providers, and the very large volumes of data, both remote sensing and model output, that are being collected, analyzed and synthesized. A related but quite distinct question also being asked by Earth scientists is, how can I apply the computations and analyses, that I commonly perform on my own data, to other data sets that may be stored on a remote computer system? Scientists are asking this question because they have powerful analytic capabilities to bring to bear on their measurements and model output.

There is technology already in place to support text search in Web-accessible archives and there is a great proliferation of Web sites that catalog information about scientific data sets or metadata. Web-based search tools have enabled relatively sophisticated interrogation of metadata collections. In parallel to these developments, the relatively slow increase in wide area network bandwidth (compared to processor speed and storage capacities) has stimulated the development of infrastructure (data transfer protocols) and tools (browser-based subsetting tools, analysis software) to support Internet access and analysis of digital data.

Therefore the technology exists to allow Earth scientists to answer the first type of questions outlined above using Web-based search tools or to the second type of questions using data transport and analysis tools. The natural next step, once a data set has been found, would be to obtain a subset of that data set or conduct some analyses of that data set in order to determine its applicability. Conversely, in order to be able to make analytic comparisons of a data set at one location with a data set at another location (that may be a data set at a backbone data center as defined by NewDISS compared, say, with a model data set at a science data center), it is necessary to know where the data sets reside. Nevertheless, the technology to answer the first type of question and then answer the second type of question, or for that matter vice versa, does not exist. This problem is a result of the fact that the two lines of technology – namely catalog, master directory (such as GCMD) and Web-based search tools, on the one hand; and data transport, data subsetting and remote interactive analysis, on the other hand – have developed independently. A scientist, who searches for and finds descriptions of data that might be useful in addressing the scientific questions he or she poses, cannot access those data directly. Likewise, a scientist, who downloads or analyzes subsets of a remotely served data set, has no way to know what analogous or similarly applicable data sets may be available, elsewhere.

SEEDS needs to develop the standards that would apply to combine the relevant technologies supporting data discovery, data ordering and data analysis. This white paper explores precisely these standards.

Standards for On-line Data Access, Analysis and Data Ordering

At first, we emphasize that we are only addressing issues related to distributed systems, most appropriate for SEEDS. Although what we are proposing here may be applicable to centralized or specialized data systems, the nature of distributed systems is particularly relevant to SEEDS.

A guiding principle ought to be *not to impose standards to the extent possible.* Rather, to adopt what is already widely used, both in open systems (Web) or science communities (most often open systems as well). Another guiding principle ought to be *science drives*, i.e. the chosen information technology solutions need to be driven by science needs, rather than vice versa. Or to put it in another way, science drivers define system requirements.

One can envisage two paradigms: In the first, the user transports data or data subsets to transfer to her own desktop for analysis. This generally may require large disks and large bandwidth if the data sets are large (as most are these days, 100's of MB for single files to GB for multiple files). The second allows distributed analysis on-the-fly at the server and transport back to the client the results of the analysis. This generally requires large data storage and easy access capabilities at the data provider (server), distribution of load for such analysis at the servers as well as a willingness of the data providers (backbone data centers, science data centers, etc.) to support such usage by users. What we are outlining below would support both paradigms.

Below, we list generalized principles that can be considered to be SEEDS standards for the problem at hand.

Data Access/Data Discovery:

- Support a wide variety of metadata with minimum semantic enforcement.
- Adopt what is already in use in various metadata for Earth system science disciplines such as atmospheres, land, oceans, coupled systems, etc.
- Adopt what promotes interoperability and community standards (i.e. open systems such as XML).

Data Analysis

- Interoperability between remote sensing (such as NASA EOS data) and model data (such as NCEP data)
- Support different data types, namely swath, gridded and point data for science users, GIS for general users, etc.
- Enable easy transport of data or subsets of data
- Format transparency (e.g. netCDF, HDF, GRIB, native form, etc.)
- Desktop tool flexibility (e.g. Matlab, IDL, GrADS, Ferret, etc.)
- Adopt formats, tools and interoperable standards that are already in use in various Earth system science disciplines such as atmospheres, land, oceans, coupled systems, etc.

Data Ordering

- If entire data sets are needed to be transported, conform or adopt what data providers already use (such as ftp, CD's, etc.).

In conclusion, in order to provide for more widespread Earth science data distribution and analysis capabilities, it is imperative that digital data access, subsetting and analysis and metadata

search be more tightly integrated. What we have proposed here has arisen from our experience and participation in the ESIP Federation and we believe very applicable to SEEDS.